# Exploratory Data Analysis on AIR BNB

**Naman Veer, Pulukuri Neil Samuel Sadanand**
**Nikhileshwar Karamala, Sarvana Adithya**
**Data science trainees,**
**AlmaBetter, Bangalore**

## 1. Introduction

Air Bnb is a company which provides lodging facilities at relatively less prices in United states of America. It was founded by three friends Brian Chesky, Joe Gebbia, Nathan Blecharczyk in the year 2021. Presently it has a revenue of 113 billion USD. The main motive for this company to start was to deal with problem of renting a place and also providing security.We have been given the data set to perfom exploratory data analysis.

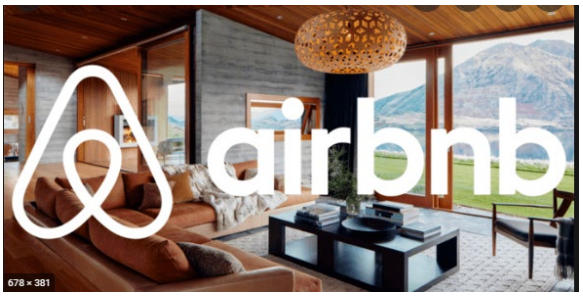



**Fig 1.0 Airbnb**

## 2. Types of Rooms

- Home/Apartment
- Private Room
- Shared rooms

Based on the dataset Air Bnb provides three kind of rooms mainly home/apartment.that can be rented on daily basis to monthly basis. Based on the requirement the customer can opt for daily or monthly usage. The second type of rooms are Private Rooms which can be quite cheaper and third rooms are Shared rooms which are the cheapest

## 3.Exploratory Data Analysis

Based on the data set given we performed exploratory data analysis and the following aspects were found.

- **Null Values**

As a part of Exploratory Data Analysis the first step is find the null vales. With null values present in the data it would result in an faulty model. The null values can be eliminated by fillna method or the replace method. Generally, the null values are filled with mean if there are different numerical values and filled with mode if the column has mostly repeated numerical values.

| | First Score | Second Score | Third Score |
|---|---|---|---|
| 0 | 100.0 | 30.0 | NaN |
| 1 | 90.0 | 45.0 | 40.0 |
| 2 | 90.0 | 56.0 | 80.0 |
| 3 | 95.0 | 56.0 | 98.0 |

**Fig 3.1 NaN images**

- **Categorical Values**

The next step is to get rid of Categorical values which have the type as object. To

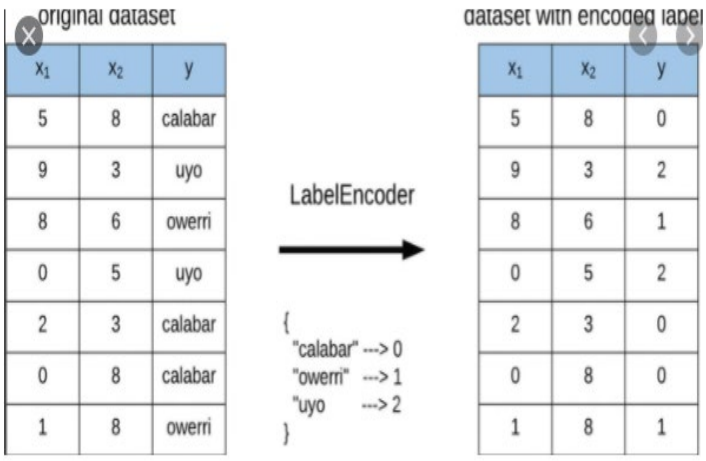
implement a machine learning model it is important we have only numerical values. Then we can train the data set and the model will be able to predict the required output. Categorical data can be converted to numerical via Dummy Encoding or Label Encoding. Dummy Encoding is used if there are less named data(sex,weather_type,etc..) whereas Label Encoder is used when we have different named data (names of a person, names of plants, names of cars ,etc. ..)

| | Person | Education | Salary(annum) |
|---|---|---|---|
| 0 | amit | Under-Graduate | $90k |
| 1 | vishal | Diploma | $80k |
| 2 | john | Under-Graduate | $90k |
| 3 | marry | Diploma | $60k |
| 4 | sherin | Under-Graduate | $90k |
| 5 | komal | Under-Graduate | $100k |
| 6 | jay | Under-Graduate | $60k |
| 7 | shree | Under-Graduate | $100k |
| 8 | kishore | Diploma | $90k |
| 9 | geetha | Diploma | $70k |
| 10 | savitha | Under-Graduate | $50k |
| 11 | vinith | Under-Graduate | $90k |

**Fig 3.1 Categorical data**

| | Person | Education | Salary(annum) | Diploma | Master's | Under-Graduate |
|---|---|---|---|---|---|---|
| 0 | amit | Under-Graduate | $90k | 0 | 0 | 1 |
| 1 | vishal | Diploma | $80k | 1 | 0 | 0 |
| 2 | Sindhu | Master's | $160k | 0 | 1 | 0 |
| 3 | Sanju | Master's | $200k | 0 | 1 | 0 |
| 4 | john | Under-Graduate | $90k | 0 | 0 | 1 |
| 5 | marry | Diploma | $60k | 1 | 0 | 0 |
| 6 | sherin | Under-Graduate | $90k | 0 | 0 | 1 |
| 7 | komal | Under-Graduate | $100k | 0 | 0 | 1 |
| 8 | jay | Under-Graduate | $60k | 0 | 0 | 1 |
| 9 | shree | Under-Graduate | $100k | 0 | 0 | 1 |
| 10 | kishore | Diploma | $90k | 1 | 0 | 0 |
| 11 | geetha | Diploma | $70k | 1 | 0 | 0 |
| 12 | savitha | Under-Graduate | $50k | 0 | 0 | 1 |
| 13 | vinith | Under-Graduate | $90k | 0 | 0 | 1 |

**Fig 3.2 One hot Encoding**




| $x_1$ | $x_2$ | y |
|---|---|---|
| 5 | 8 | calabar |
| 9 | 3 | uyo |
| 8 | 6 | owerri |
| 0 | 5 | uyo |
| 2 | 3 | calabar |
| 0 | 8 | calabar |
| 1 | 8 | owerri |

| $x_1$ | $x_2$ | y |
|---|---|---|
| 5 | 8 | 0 |
| 9 | 3 | 2 |
| 8 | 6 | 1 |
| 0 | 5 | 2 |
| 2 | 3 | 0 |
| 0 | 8 | 0 |
| 1 | 8 | 1 |

**Fig 3.3 Label Encoding**

- **Description of data**

After Converting the dataset into numerical data then the next step is to get the overall description of data set. This includes function like info , describe ,dtypes . The info method gives the column types as well as the number of count column. The describe method gives the various statistical information like mean , median ,mode etc.. The dtypes method gives us the data type of each column.

```
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
```

**Fig 3.4 Output of info method**

- **Data Cleaning**

  The next stage is very important as we have to select the appropriate columns. We cannot process the entire dataset for couple of reasons. First if we load the entire data set then the computation type would be more and we would deal with vague information. So it is important to get rid of this vague data. We can use drop method by which we can delete multiple columns. We can delete via index or via labels. We can use iloc for index and loc for labels. .

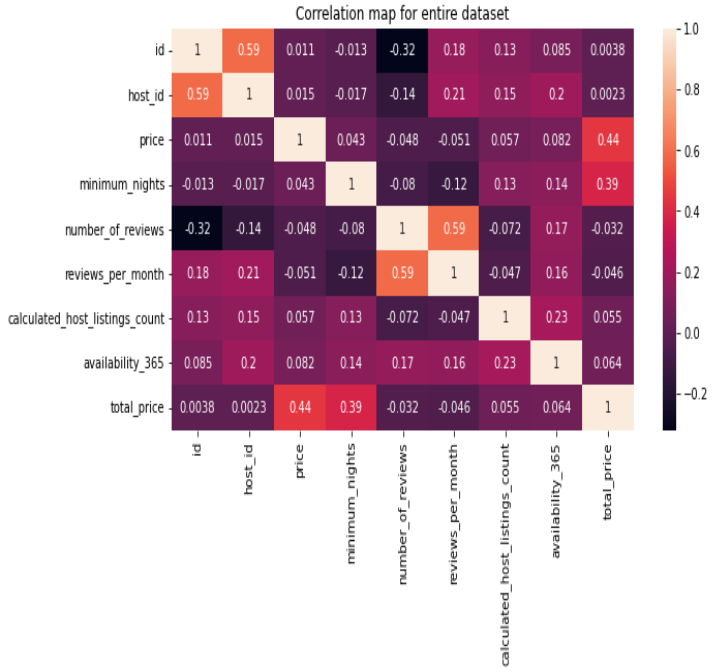
- **DATA Analysis**

  The next step is to perform data analysis. We can group the data by group by operations to explore the data. We can come to various conclusions or trends based upon analysis. This step is important because we became familiar with data set and if the client asks for some insights we will be in a position to provide him/her the insights. In our data set we performed multiple analysis starting from analyzing the prices , the reviews, the type of rooms and the localities where the properties are expensive. Furthermore we were able to also provide correlation between different parameter.

- **DATA VISUALIZTION**

  This analyzed data should be presented in the form visuals as business holders or clients prefer visualization rather than observing large chunks of results. We can take the help of matplot lib and seaborn to perform data visualization. The first library provides a great range of plots starting from count plots, bar plots, box plots, histograms and many more. The latter which is seaborn is very good in terms of visualization. Complex data sets can be represented easily via seaborn . Many prefer seaborn over matplotlib .We can also plot athree way plot via seaborn .The most important analysis is the correlation which can easily be plotted via seaborn.

**Fig 3.5 Correlation Seaborn Plot**




## 4.Conclusion:

Finally we have finished with are exercise of exploring a data set which is the first step toward building a machine learning model. As a part of exercise we see that we had to go

through eliminating the null values, converting the categorical data into numerical data , getting info and then cleaning the data, and then finally analyzing the data using visual tools like seaborn and matplotlib

**References-**


1. Machine Learning by W3 schools
2. Machine Learning by Krish Naik
3. Pandas by geek for geeks